Biased Restaurant Reviews: Starbucks in NYC

Hsuan-Hsueh Huang *[1], Jiangyue Mao †[2], and Yuqi Yan‡[3]

[1,2,3]Department of Statistics, University of Michigan

1 Introduction

1.1 Research Question

In recent years, online customer reviews or product reviews have become the major source for collecting feedback from customers. Restaurant owners could find ways to improve their service or even evaluate whether a new store is easy/hard to gain good ratings and reviews in a specific location by analyzing those online reviews. In this project, we want to explore the factors that cause regional bias in restaurants' ratings. We chose Starbucks as the target store to analyze since it is our favorite and the most popular chain store in U.S.A. We are curious about why the same brand with almost the same products could receive different ratings in different locations. The analysis will first focus on boroughs in New York City due to the time limit and the project scope. By analyzing the online review text with regional data (i.e. income levels) of Starbucks in different regions, we want to answer the following two questions:

1. What factors do customers care about Starbucks?
2. If those factors varied by region? If so, do income levels make them different?

The analysis result of this case study will provide the business owners the insight into how to satisfy the customers and run a business successfully in each region.

*hsuanhs@umich.edu
†maojy@umich.edu
‡yukei@umich.edu

1.2 Related Work

There are several works [1][2][3] have studied the potential indicator of the success of restaurants. Lian et al. proposed four factors, geography, user mobility, user rating, and review text, to predict the long-term survival of restaurants.[1] They found that the location and nearby places, user mobility, and review text, are important factors, while consumers' ratings or sentiment is hard to provide enough insights. A large-scale randomized experiment on a social news aggregation Web site was proposed by Lev et al. to investigate whether aggregated digitized opinions of others distort decision-making, and they concluded that prior ratings created a significant bias in individual rating behavior.[2] James collected and analyzed data from Yelp to answer two proposed questions, "Is the service better in restaurants in higher-income areas than in lower-income areas?" and "Is there a correlation between income levels and restaurant ratings?", and the results showed that the income of a neighborhood, population density, and the number of reviews a restaurant receives have nothing to do with the ratings of its restaurants.[3] In addition to the factors of success of restaurants, some researchers study the review text and attempt to apply natural language processing techniques to analyze such text data. A combined CNN-LSTM architecture was proposed to analyze the sentiment of restaurants reviews and got an accuracy of 94.22%.[4] Heng et al. proposed the LCF-ATEPC, which is a multi-task learning model that solves aspect term extraction (ATE) and aspect polarity classification (APC) simultaneously.[5] Dan et al. investigate linguistic structure in 900,000 online restaurant reviews to explore the narratives that consumers use to frame positive and negative sentiment.[6]

2 Data

We collected and combined two type of data, online review data and regional data, in our analysis for finding the relationship between them and the rating of Starbucks in different regions.

2.1 Online Review Data

We decide to use Google Maps Review as the source of the online review data. Crawling such review data from Google Map Review is not an easy task since its public API limits the retrieved results to the latest five. Also, Google Maps Review applies Ajax to update the content of the web page, which means the crawler has to simulate a user that keeps scrolling down to get all the review data. We finally found a commercial web scraping tool, Octoparse,[7] that satisfied

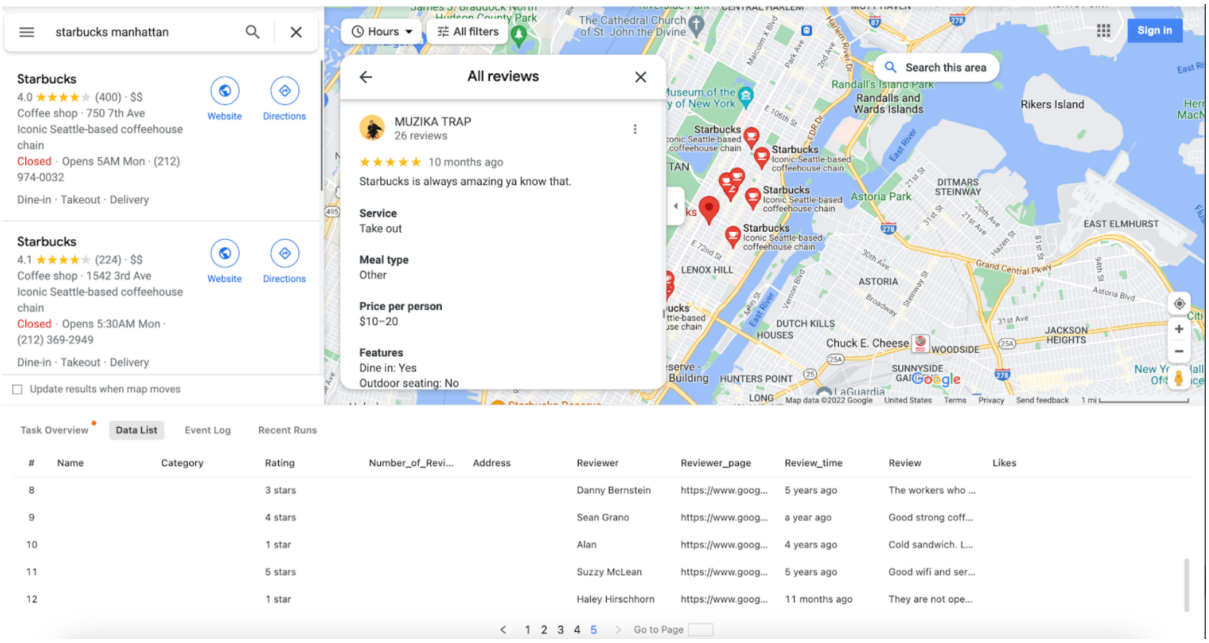

Figure 1: Example of Octoparse

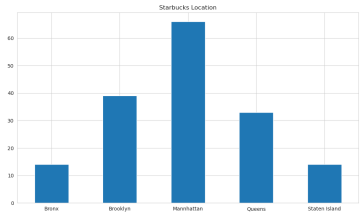

Figure 2: Count of Starbucks stores

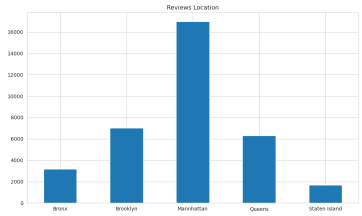

Figure 3: Count of Starbucks reviews

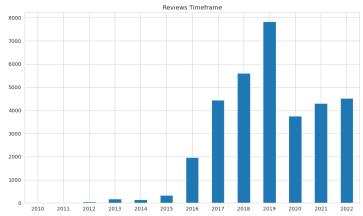

Figure 4: Timeframe of Starbucks reviews

our requirements. Figure 1 shows a use case of operating Octoparse. Given the URL of the store reviews in Google Maps, it can automatically collect the rating, review text, review time, and the number of likes of each review for our analysis. Figure 2 3 4 show the distribution of the data.

2.2 Regional Data

In order to find the relationship between Starbucks ratings and household income, we divided NYC household income levels into two by boroughs: boroughs with dark blue indicate there are greater or equal to 10% of the households with an annual income of greater than $20,0000 (high income) which is consisted of Manhattan, Brooklyn, and Staten Island, and boroughs with yellow mean there are less than 10% of the households with an annual income of greater than

$20,0000 (low income), which is consisted of Bronx and Queens[1]. We only divided the household income levels into two instead of more bins because it can help clarify the rating and review differences across different regions. If we look into the income levels more closely, we will see that lower Manhattan and downtown Brooklyn are the regions with the highest household income.

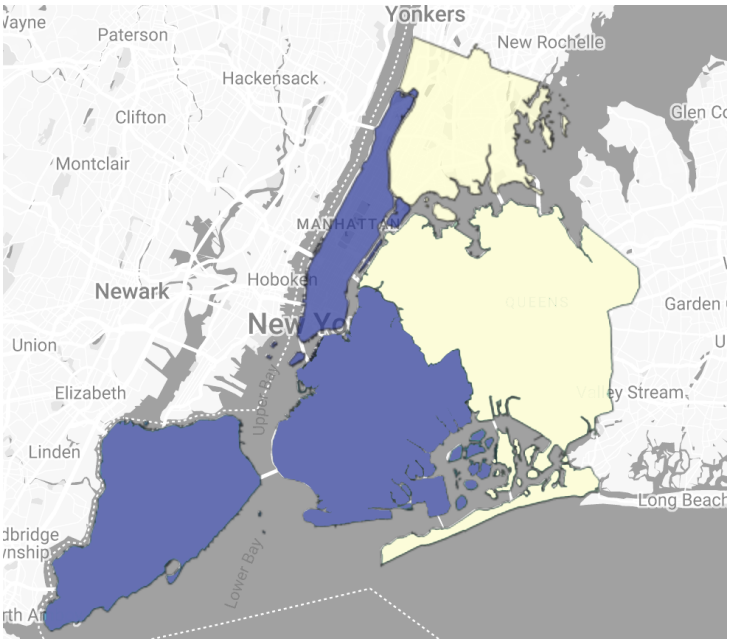

Figure 5: NYC household income

3 Methods

In this section, we'll discuss the methods we used to process the users' review data. To understand the customers' reviews and the related sentiment towards certain aspects of Starbucks, we implemented two models:

- Aspect Extractor: extract the opinion aspects and polarity from each review sentence
- Aspect Classifier: classify the aspects for further analysis

Figure 6 shows the overview aspect-extraction and classification pipeline used in the study. Section 3.1 will introduce the preprocessing methods we used to transform the wild Google Map data. The implementation of these two models is discussed in section 3.2 and section 3.3, and the correlation analysis is discussed in section 3.4.

[1]Source: https://data.cccnewyork.org/data/map/29/household-income#29/34/2/52/62/a/a

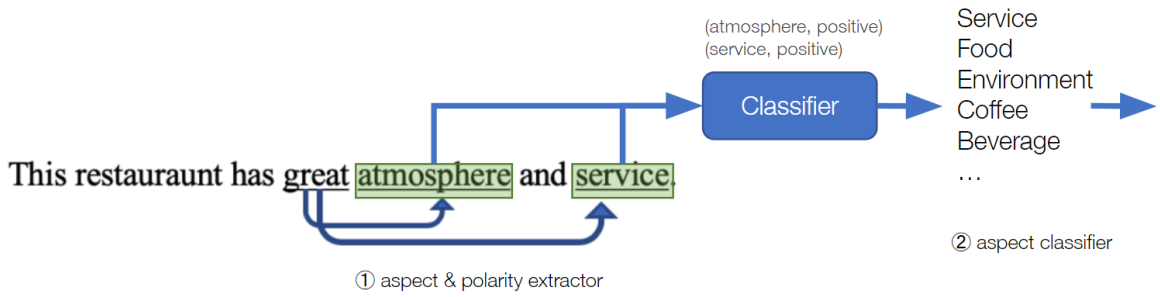

Figure 6: Sentiment analysis pipeline of users' review

3.1 Data Preprocessing

Multilingual reviews: Reviews from Google Maps are from all over the world and we have reviews in different languages. We used English reviews and the translated results produced by Google Translate for other languages because our study focuses on Starbucks in the U.S. We're assuming most of the reviews from NYC are in English and our model is trained for the English language specifically.

Text processing: Emojis are removed because of lacking training dataset of reviews data with emojis. The reviews are split into short paragraphs (less than 200 words for each short paragraph) due to the limitation of input size. Details will be discussed in 3.2. The aspects words are then processed with WordNet[2] Lemmatizer to reduce the vocabulary size. WordNet is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept.[8]

3.2 Aspect Extraction

Aspect extraction, in other words, is to label each word of the sentence to be "aspect words" or "not aspect words". This is also known as sequence labeling in NLP. A standard method for sequence tagging is inside-outside-beginning (IOB). As depicted in Figure 7, "B" labels the beginning word of an aspect word, and "I" is for non-beginning aspect words. Non-aspect words are labeled with "O". e.g. the input of the review "The food price is reasonable here." will be first tokenized as $S = \{w_1, w_2, ..., w_n\}$, where w stands for a token and $n = 7$ is the total number of tokens. The sentence will be labeled as $Y = \{O, B_{asp}, I_{asp}, O, O, O, O\}$.

Aspect polarity extraction is a subtask of sentiment analysis. The goal is to predict the aspect

[2] https://wordnet.princeton.edu/

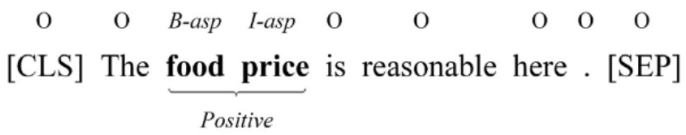

Figure 7: The IOB format review input of BERT

polarity for the aspects words. e.g. for the review input "The food price is reasonable here." and its tokenized result $S = \{w_1, w_2, ..., w_n\}$, the label will be $Y = \{2, 3, POS\}$.

For this study, we're using the labeled restaurant reviews training data from SemEval-14[3] because it's domain-specific and the restaurant reviews are similar to Starbucks reviews. SemEval-14 provides both IOB and the polarity labels for each aspect word. Figure 8 shows some samples from the SemEval-14 dataset.

No.	Sentence	Aspect	Polarity
1	Great laptop that offers many great **features**!	**features**	positive
2	The **seats** are uncomfortable if you are sitting against the wall on the wooden benches.	**seats**	negative
3	How do you settlers of catan for the **xbox**?	**xbox**	neural

Figure 8: Several samples from SemEval-14 dataset.

To train an aspect extractor for both aspect words and aspect polarity, we used a multi-task learning network ATEPC[5] (Aspect Term Extraction and Aspect Polarity Classification). As shown in Figure 9, ATEPC is a fusion network for both local context and global context. The identification of local context depends on the semantic-relative distance (SRD). SRD describes how far a token is from a targeted aspect. SRD is calculated as:

$$SRD_i = |i - p_a| - \lfloor \frac{m}{2} \rfloor \quad (1)$$

where $i(1 < i < n)$ is the position of the token, p_a is the central position of each aspect term, m is the length of the aspect term, and SRD_i denotes the SRD between the ith token and the aspect term.

We use ATEPC in our study because it supports both aspect extraction and polarity extraction at the same time, and the fusion of local and global context helps to extract aspects from long reviews we have from Google Maps. The model is trained with PyABSA[9] and PyTorch. Due to

[3] https://alt.qcri.org/semeval2014/task4

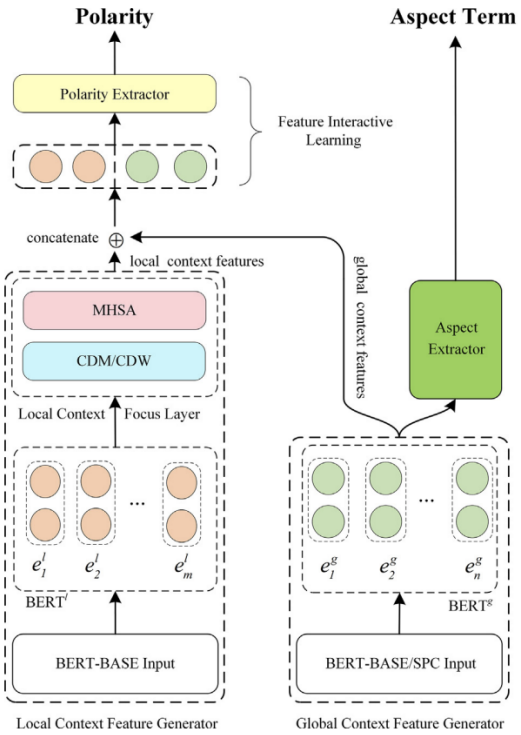

Figure 9: Multi-task learning framework of ATEPC[5]

the limitation of time and computing resources, we limit the input sentence size to at most 200 tokens. The network reached an F1 score of 0.813 in aspect extraction, and an F1 score of 0.865 in aspect polarity extraction with the SemEval-14 restaurant dataset. Table 1 shows some sample Starbucks reviews in NYC and the results we get from the model.

Review Text	**Aspects**	**Polarities**
I ordered an iced pumpkin spice latte with cold foam, and they gave me an iced latte. it's not even a little bit orange, or sweet. I paid 9.50 for it and they didn't even give me the actual drink I ordered. I didn't notice until I got home to drink it, and it's too far to go back and ask for a new one but goodness was I disappointed.	[iced pumpkin spice latte with cold foam, iced latte, drink]	[NEG, NEG, NEG]
large seating area. fast service. iconic location. (the first Starbucks in manhattan!)	[seating area, service, location]	[POS, POS, POS]
one of the oldest Starbucks in the city. small amount of wall outlets and tons of unwelcome crowds at night but it's a nice decent place to drink coffee and relax. restroom always has a long line.	[outlets, crowds, place, coffee, restroom, line]	[NEG, NEG, POS, POS, NEG, NEG]

Table 1: Startbucks review samples from NYC and the aspect extraction results

3.3 Aspect Classification

Word Embeddings: to train the classification model, we extract the features out of the aspect words. We used the pretrained sentence-transformers/all-MiniLM-L6-v[4] from HuggingFace to map the aspect words to a 384-dimension dense vector space.

Mini-batch K-means is a variant of the popular K-means clustering algorithm. Figure 10 shows the pseudo code for the algorithm. Instead of generating centroids for the whole dataset, Mini-batch K-means randomly picked a subset at each iteration to reduce the computation cost. The clustering results are visualized using tSNE and Word cloud. As shown in Figure 11, we initially set $K = 10$ clusters. Then we manually merged the clusters with similar semantics. We'll discuss the final results in section 4.2.

Algorithm 1 Mini-batch k-Means.

Given: k, mini-batch size b, iterations t, data set X
Initialize each $\mathbf{c} \in C$ with an $\mathbf{x}$ picked randomly from X
$\mathbf{v} \leftarrow 0$
for $i = 1$ to t **do**
  $M \leftarrow b$ examples picked randomly from X
  for $\mathbf{x} \in M$ **do**
    $\mathbf{d}[\mathbf{x}] \leftarrow f(C, \mathbf{x})$ // Cache the center nearest to $\mathbf{x}$
  end for
  for $\mathbf{x} \in M$ **do**
    $\mathbf{c} \leftarrow \mathbf{d}[\mathbf{x}]$ // Get cached center for this $\mathbf{x}$
    $\mathbf{v}[\mathbf{c}] \leftarrow \mathbf{v}[\mathbf{c}] + 1$ // Update per-center counts
    $\eta \leftarrow \frac{1}{\mathbf{v}[\mathbf{c}]}$ // Get per-center learning rate
    $\mathbf{c} \leftarrow (1 - \eta)\mathbf{c} + \eta\mathbf{x}$ // Take gradient step
  end for
end for

Figure 10: Mini-batch K-means algorithm[10]

3.4 Correlation Analysis

Spearman's Rank Correlation is a popular nonparametric metric for measuring the strength and direction of association between two ranked variables. i.e. how well the relationship between two variables could be represented using a monotonic function. The Spearman's rank correlation coefficient is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

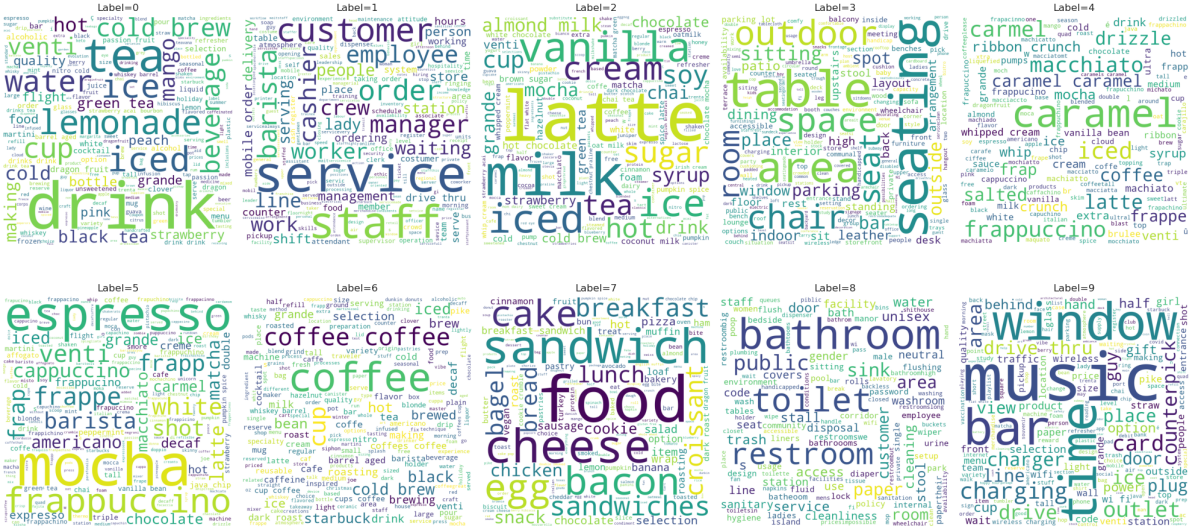

Figure 11: Word cloud of clustering results (K=10)

where d_i is the difference between the two ranks of each observation and n is the number of observations. ρ takes values from -1 to +1. $\rho = 1$ means a perfect association of rank.

With pairwise Spearman's rank correlation coefficient, we can find possible related factors of different sentiments towards different aspects of the restaurant.

4 Findings

4.1 Ratings

Figure 12 indicates that most Starbucks in NYC are distributed around lower Manhattan and downtown Brooklyn, which makes sense since they are the highest income regions. Boroughs with high income levels received more than 20,000 ratings and boroughs with low income levels received less than 10,000 ratings. We visualized the mean ratings across all Starbucks in NYC using PowerBI and our findings are in Figure 13.[5] The lighter the points are, the lower the mean ratings are. It can be seen that the mean ratings in areas with relatively high household incomes are lower than that of other places. In addition, the mean rating of Starbucks in Brooklyn, Manhattan, and Staten Island is 3.68, whereas the mean rating of Starbucks in Queens and Bronx is 3.83, which is around 4% higher than that of the mean rating of high-income

[5]There are over 150 Starbucks in NYC, in order to display all of the points clearly, we zoomed in and cropped the visualization, and made a collage from them. There might be some inconsistencies regarding the specific locations on the image. The detailed ratings for each borough are in the appendix

neighborhoods.

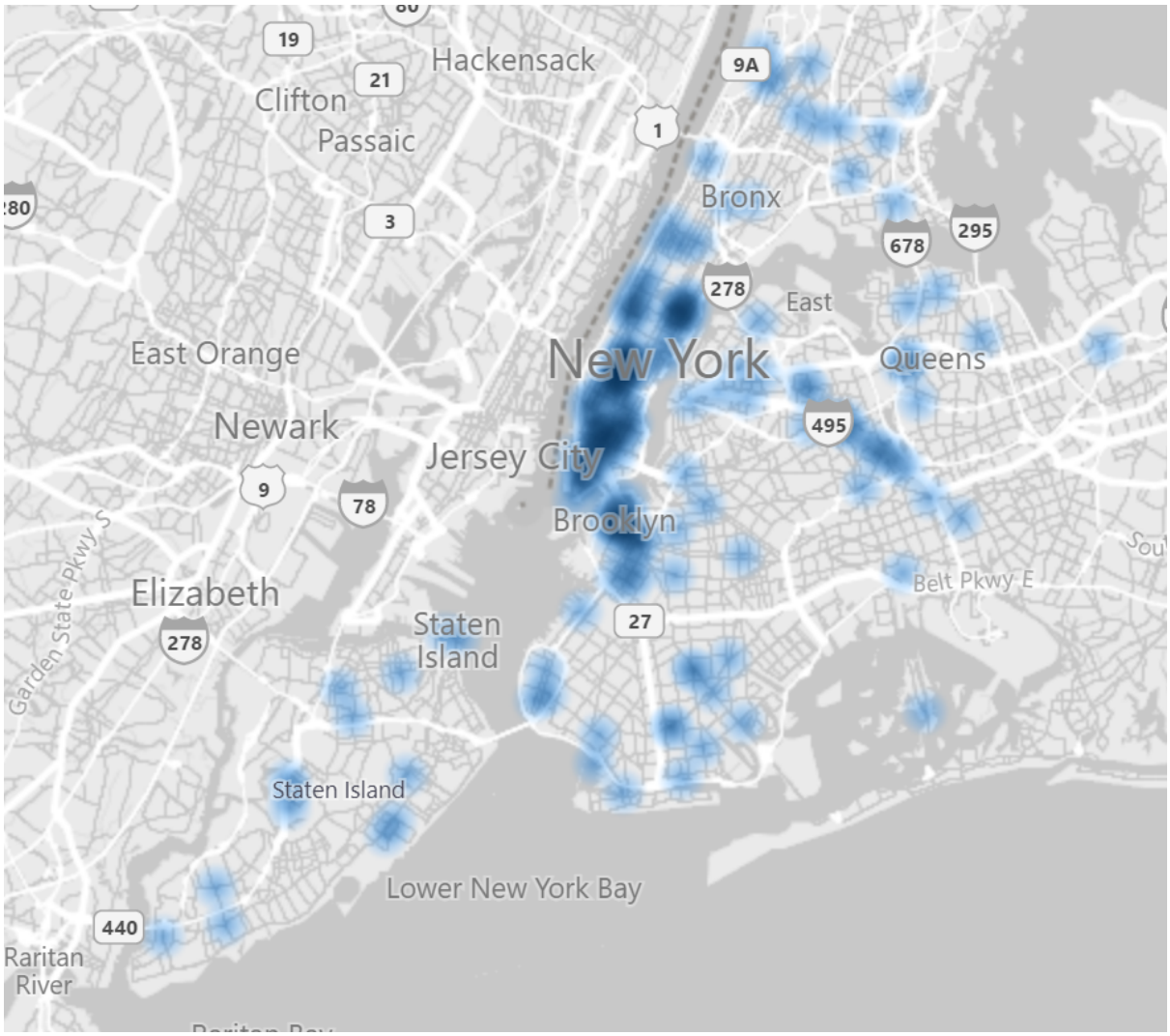

Figure 12: Overall Starbucks distribution in NYC

4.2 Reviews

Based on our correlation analysis, we found out that the reviews from regions with high income levels are mostly correlated with waiting time with a factor of 0.6 whereas reviews from regions with low income levels are mostly related to beverages and environment with a factor of around 0.55. However, since correlation does not mean causation, we performed sentiment analysis and aspect embedding clustering to analyze the reviews more rigorously.

We first analyzed the overall reviews across all regions. By setting k = 10 for aspect embedding clusters and merging similar clusters, we have six clusters showing six keywords as Figure 14, which are “service”, “coffee”, “wait”, “env”, “food”, and “beverage”. We can see that for “service”,

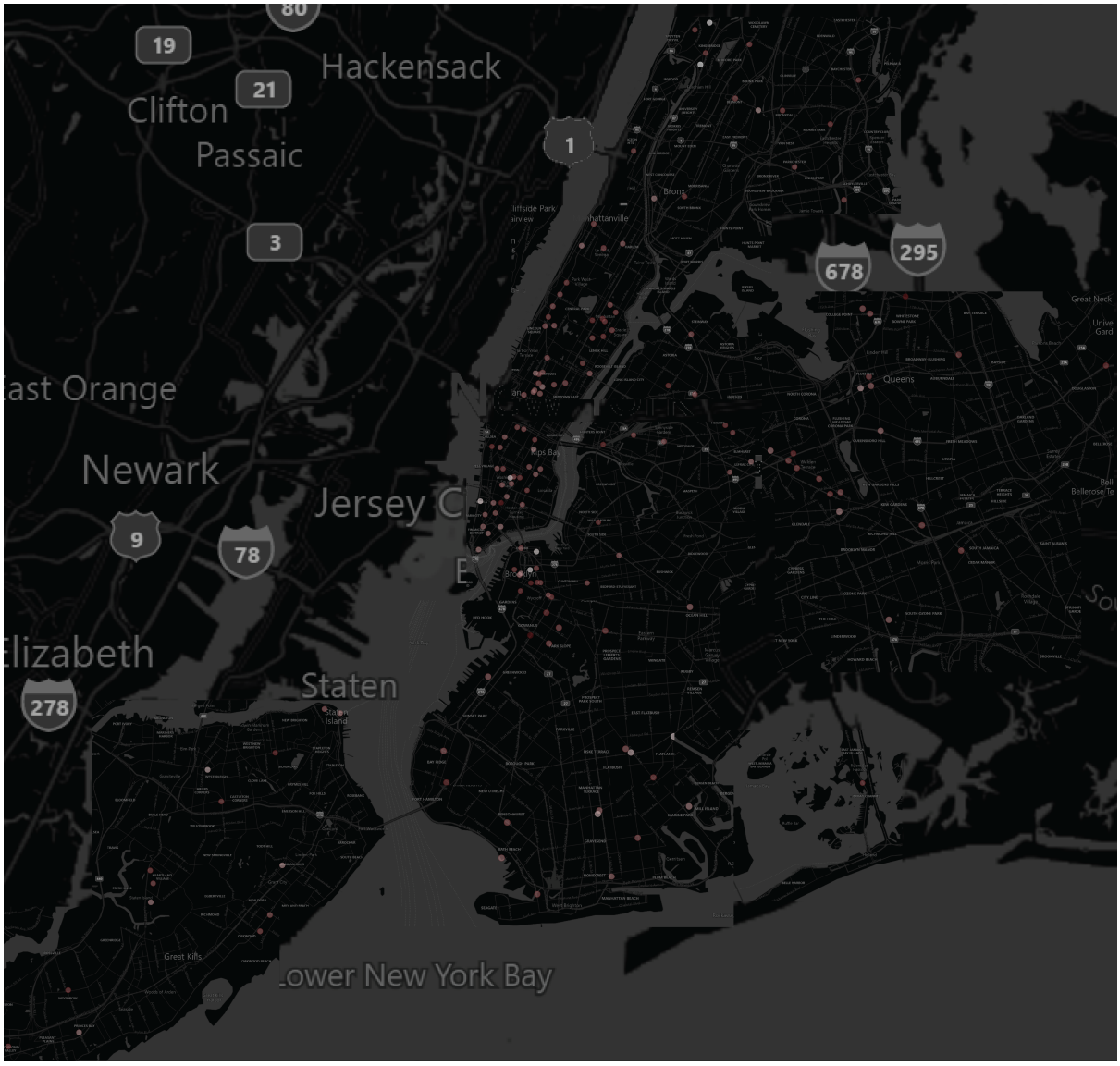

Figure 13: Starbucks ratings in NYC

the customers mainly mentioned “staff” and “barista”. “latte”, “ice”, and “frappuccino” play an important role in reviews related to “coffee”. “music” and “drive (through)” are most related to reviews with “wait”, and “table” and “seating area” are most related to reviews with “env”. “tea” and “drink” are closely related to reviews with “beverage”. Those reviews indicate that the ratings and reviews are not only based on the products that Starbucks offers, but also on the service and environment that Starbucks has. Due to time constraints, we did not perform clustering per income level, which could be done given more time.

Figure 15 shows the positive and negative reviews proportion for each region. There is not a significant difference between positive and negative reviews across different regions regarding

Figure 14: Aspect embedding clusters

review aspects. Based on the positive reviews, we can see that customers across all regions are more satisfied with service and coffee by around 20% to 60% compared with waiting time, food, and environment. Interestingly, service also contributes most to the negative reviews across all regions together with waiting time. The customers are less satisfied with them by around 10% - 60% compared with other aspects. In addition, customers in Queens seem to be more satisfied with the environment, and customers in the Bronx seem to be more satisfied with service, compared with customers who wrote reviews in other regions.

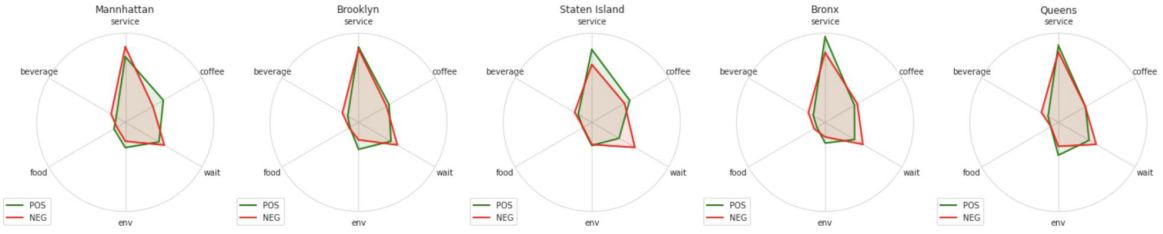

Figure 15: Positive and negative reviews proportion per region

In order to compare review differences more clearly, we have Figure 16 showing the positive vs. negative reviews ratio. We can see that the ratios of all aspects are similar across all regions,

with the positive reviews from high income regions leaning towards coffee more, and the positive vs. negative reviews ratio related to food being around 25% higher in Manhattan compared with that in other areas. The ratio related to service, beverage, and waiting time across all regions is close.

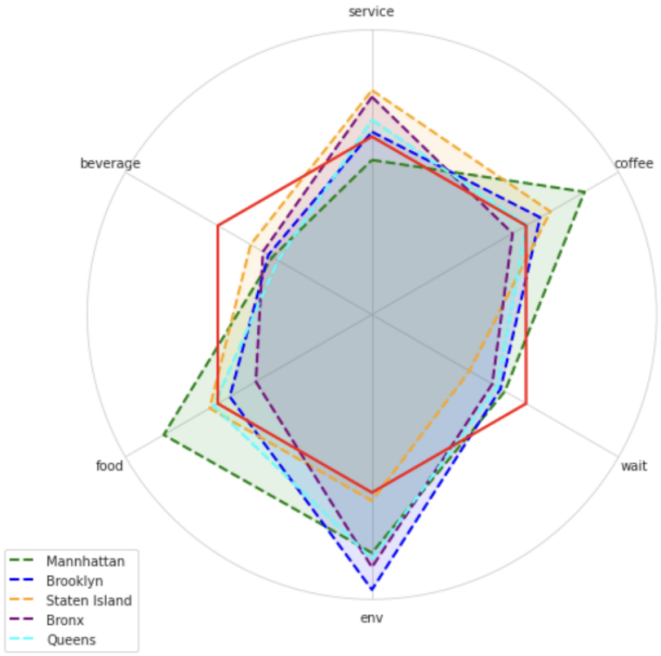

Figure 16: Positive/negative reviews proportion ratio across all regions

5 Discussion

In this project, we want to understand customers' reviews from Starbucks in NYC and figure out whether the reviews are biased in different stores. If so, we want to find out what's the differences and find the relationship between the reviews and income levels. We crawl the Starbucks reviews data from Google Maps. The reviews are first cleaned and processed. Then the aspect words are extracted and classified using the ATEPC network and K-means clustering.

Overall, for NYC specifically, although the ratings of Starbucks are only 4% higher in low income regions, according to our literacutre review, this might be because business owners might spot an opportunity to gain customers by improving their customer service, especially in low income areas. In addition, patrons living in low income neighborhoods would have the ammunition to demand better customer service in the restaurants they frequent. And city hall officials might also see this as a form of income segregation and work harder to ensure restaurants in low income areas have the same amenities and service levels as those in middle or high income

areas. However, we did not see a strong relationship between reviews and income levels.

5.1 Limitations

- We've made an assumption that customers going to Starbucks have the same demographic distribution as the local citizens. However, this assumption doesn't necessarily hold. The main target segment of Starbucks customers aged from 25 to 40 with high incomes, and the second target group is 18 to 24 years of age and belongs to richer families.[11] We made this assumption because the actual customers' information is unavailable due to privacy policy. Real customers' information will give more reasonable results in the correlation analysis.
- The aspect extractor is trained on the SemEval-14 restaurant review dataset. The distribution may differ from our dataset. The accuracy of the aspect extractor will largely affect the ad hoc analysis.
- The aspect classification is done by using unsupervised methods. We still need to manually identify the meaning of each cluster. We also don't have control over the level of granularity that how these aspects are classified.
- Due to the limited availability of demographic information, we mainly focus on Starbucks near NYC. The findings may or may not be applied to other coffee shops or regions.
- The data points are not sufficient to understand customers' preferences in different areas. General aspect words take up to 40% of our categories, e.g. "coffee" for the coffee category.

5.2 Future Work

- We could try to apply our methods to a broader region with more data points such as some states in the US.
- If we have more time we could separate the reviews by high and low income regions and perform clustering analysis on it.
- For future study in specific aspect categories, we may need to figure out what categories we should use, label the aspects and use supervised methods instead.

6 Contribution

All co-authors discussed the research questions and were involved in writing the report.
Yuqi Yan was responsible for building the models and designing the experiments. She transformed the data and tested several language models for aspect extraction and aspect classification. Then she conducted the ad hoc sentiment analysis.
Jiangyue Mao was in charge of visualizing the results. She also summarized all findings.
Hsuan-Hsueh Huang was responsible for building our own dataset, including scraping data from Google Maps Review and labeling the regions.

7 Acknowledgement

The completion of this project could not have been accomplished without the support of all our classmates from SI699, who gave us lots of rewarding feedback and suggestions. Also, we want to thank Professor Justine Zhang for leading this course and holding the one-on-one weekly meetings to answer our questions. Her guidance and advice carried us through all the stages of finishing this project.

References

[1] Lian, J.; Zhang, F.; Xie, X.; Sun, G. *Proceedings of the 26th International Conference on World Wide Web Companion*; WWW '17 Companion; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, CHE, 2017; p 993–1002.

[2] Muchnik, L.; Aral, S.; Taylor, S. Social Influence Bias: A Randomized Experiment. *Science (New York, N.Y.)* **2013**, *341*, 647–51.

[3] Mbuthia, J. Is there a correlation between a restaurants ratings and the income levels of a neighborhood? (August 2019); https://medium.com/swlh/is-there-a-correlation-between-a-restaurants-ratings-and-the-income-levels-of-a-neighborhood-5fe41165e4f1.

[4] Hossain, N.; Bhuiyan, M. R.; Tumpa, Z. N.; Hossain, S. A. Sentiment Analysis of Restaurant

Reviews using Combined CNN-LSTM. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020; pp 1–5.

[5] Yang, H.; Zeng, B.; Yang, J.; Song, Y.; Xu, R. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing* **2021**, *419*, 344–356.

[6] Jurafsky, D.; Chahuneau, V.; Routledge, B.; Smith, N. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* **2014**, *19*.

[7] Octoparse. https://www.octoparse.com/.

[8] Miller, G. A. WordNet: a lexical database for English. *Communications of the ACM* **1995**, *38*, 39–41.

[9] Yang, H.; Li, K. PyABSA: Open Framework for Aspect-based Sentiment Analysis. 2022; https://arxiv.org/abs/2208.01368.

[10] Sculley, D. Web-scale k-means clustering. Proceedings of the 19th international conference on World wide web. 2010; pp 1177–1178.

[11] Haskova, K., et al. Starbucks marketing analysis. *CRIS-Bulletin of the Centre for Research and Interdisciplinary Study* **2015**, *1*, 11–29.

A Appendix

The followings are the detailed visualizations of the ratings in the five boroughs in NYC. We selected one Starbucks in each borough to show the exact rating as a reference point.

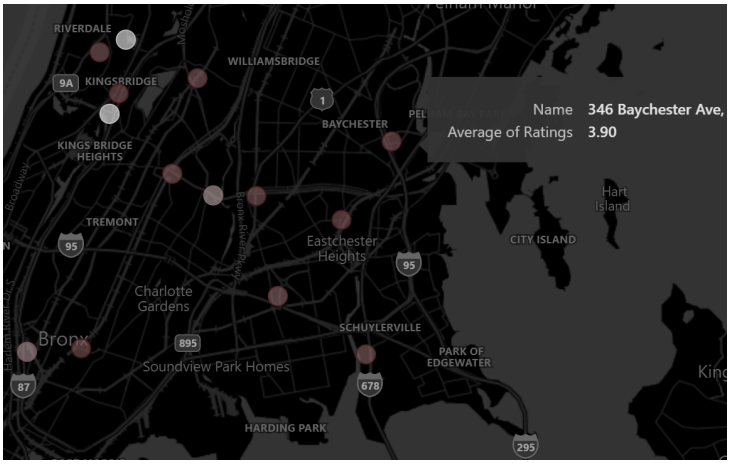

Figure 17: Bronx Starbucks ratings

Figure 18: Brooklyn Starbucks ratings

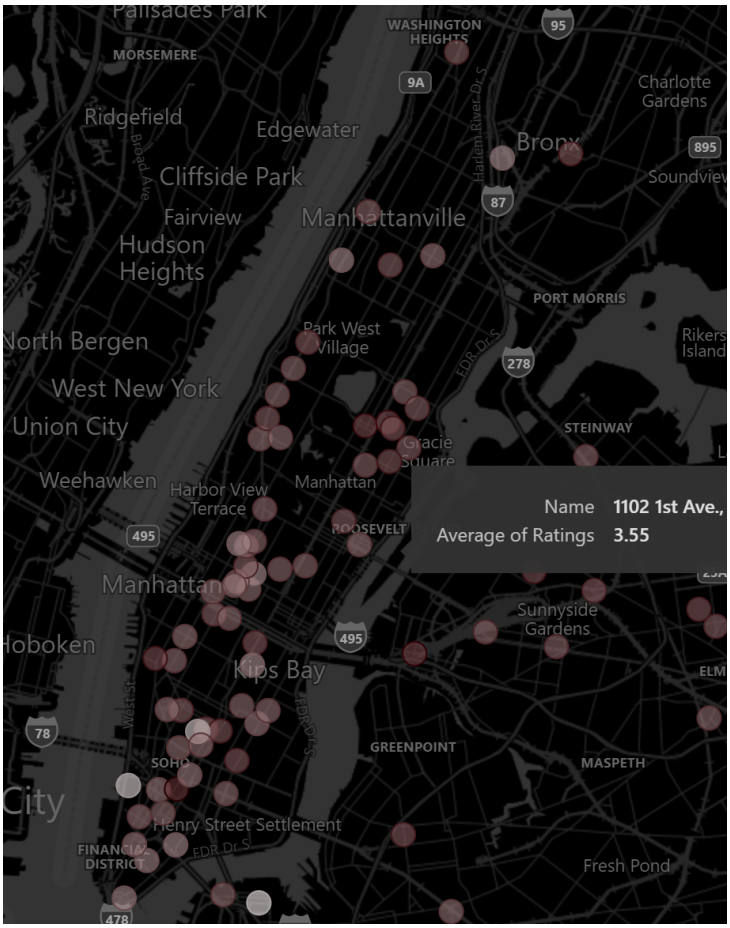

Figure 19: Manhattan Starbucks ratings

Figure 20: Queens Starbucks ratings

Figure 21: Staten Island Starbucks ratings